CLAIMS

I claim:

1. A system for automatically determining a language of a document from a set of candidate of languages, the system comprising:

logic for setting a negative assumption value indicating the document is not one of the candidate languages;

an extractor for extracting a character string from the document; and

a language analyzer for determining a probability value that the character string does not belong to the candidate languages and includes logic for adjusting the negative assumption value based on the probability value, the language analyzer determining that the document is one language of the candidate languages when the negative assumption value passes a threshold value.

2. The system as set forth in claim 1 further including probability data containing probabilities that a selected character string belongs to each of the candidate languages.

3. The system as set forth in claim 2 further including logic for retrieving the probability value from the probability data that corresponds to the character string.

4. The system as set forth in claim 1 further including an information retrieval engine for retrieving documents in response to a search request, the documents retrieved being analyzed by the language analyzer.

5. The system as set forth in claim 1 wherein the logic for adjusting includes logic for combining the negative assumption value with the probability value.

6. The system as set forth in claim 1 wherein the language analyzer further includes iteration logic for causing the extractor to extract another character string from the document if the negative assumption value fails to pass the threshold value.

7. A method of determining a language of a document from a set of candidate languages, the method comprising the steps of:

setting a null hypothesis to a true value for each candidate language indicating the document is not in the candidate language and setting a false value;

extracting a text string from the document;

determining a contrary probability for each candidate language that the text string does not belong to the candidate language;

adjusting the null hypothesis for each candidate language with the contrary probability corresponding to the candidate language; and

determining the document is one language from the candidate languages when the null hypothesis for the one language is disproved by approaching the false value.

8. The method as set forth in claim 7 further includes setting a threshold value indicating that the null hypothesis is disproved.

9. The method as set forth in claim 8 further includes repeating the extracting step for a different text string from the document and repeating the method until the null hypothesis is disproved for one of the candidate languages by passing the threshold value.

10. The method as set forth in claim 7 further includes pregenerating probability data corresponding to each candidate language, the probability data including a probability value for a text string that is normalized based on an occurrence probability of the text string in all the candidate languages.

11. The method as set forth in claim 7 further includes identifying the document based on a search request.

12. The method as set forth in claim 7 wherein the extracting step includes extracting a plurality of sequential characters that form the text string.

13. The method as set forth in claim 7 wherein the setting step includes setting the true value to 1 and setting the false value to 0.

14. The method as set forth in claim 7 wherein the contrary probability for a first candidate language is determined based on a number of occurrences of the text string found in a sample set of documents from the first candidate language which is normalized by a sum of occurrences of the text string found in a sample set of documents from all the candidate languages.

15. A process of determining that a document is in a selected language, the process comprising the steps of:

setting a probability assumption indicating that the document in not in the selected language;

extracting a character string from the document; and

disproving the probability assumption based on a contrary probability that the character string does not belong to the selected language such that if the contrary probability fails to support the probability assumption, then the document is determined as being in the selected language.

16. The process as set forth in claim 15 further includes determining the document is the selected language from a set of candidate languages.

17. The process as set forth in claim 16 further including generating a probability database having a contrary probability for each of a plurality of character strings for each of the candidate languages, where the contrary probability of a character string in one language is determined based on an occurrence frequency of the character string in the one language

influenced by a total occurrence frequency of the character string in all the candidate languages.

18. The process as set forth in claim 17 further including determining the occurrence frequency of each character string based on a sample set of documents provided for each of the candidate languages.

19. The process as set forth in claim 17 wherein the contrary probability of the character string in one language is normalized by the total occurrence frequency of the character string in all the candidate languages.

20. The process as set forth in claim 15 further including identifying the document in response to a search request.